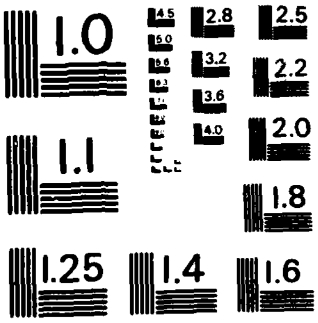
AD-A133 259 IDENTIFYING DIFFERENT ITEM RESPONSE CURVES(U) MCFANN GRAY AND ASSOCIATES INC SAN ANTONIO TX M V LEVINE SEP 83 AFHRL-TP-83-33 F41689-81-C-0012

1/1

UNCLASSIFIED

F/G 5/10 NL

END
DATE
FILMED
[illegible]
DTIC


1.0
4.5
5.0
5.6
6.3
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

AFHRL-TP-83-33

AIR FORCE

HUMAN RESOURCES

IDENTIFYING DIFFERENT ITEM RESPONSE CURVES

By

Michael V. Levine

University of Illinois
Urbana-Champaign, Illinois 61820

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235

September 1983

Interim Report for Period October 1981 – January 1982

Approved for public release; distribution unlimited.

LABORATORY

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.

JANOS B. KOPLYAY
Contract Monitor

NANCY GUINN, Technical Director
Manpower and Personnel Division

J. P. AMOR, Lt Colonel, USAF
Chief, Manpower and Personnel Division

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
AFHRL-TP-83-33

2. GOVT ACCESSION NO.
AD-A133 259

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
IDENTIFYING DIFFERENT ITEM RESPONSE CURVES

5. TYPE OF REPORT & PERIOD COVERED
Interim
October 1981 – January 1982

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Michael V. Levine

8. CONTRACT OR GRANT NUMBER(s)
F41689-81-C-0012

9. PERFORMING ORGANIZATION NAME AND ADDRESS
McFann-Gray & Associates, Inc.
5825 Callaghan Road, Suite 225
San Antonio, Texas 78228

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
61102F
2313T137

11. CONTROLLING OFFICE NAME AND ADDRESS
HQ Air Force Human Resources Laboratory (AFSC)
Brooks Air Force Base, Texas 78235

12. REPORT DATE
September 1983

13. NUMBER OF PAGES
36

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS (of this report)
Unclassified

15.a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of this abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

computer adaptive testing
item bias
item response theory
latent trait theory
logistic models
maximum likelihood estimation
test bias
test compromise
test construction
tests and measurement

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A method is presented for determining whether or not two three-parameter item characteristic curves differ significantly from each other. The method may be used within the context of item response theory to detect evidence of item compromise, change with time, or group-specific differences (e.g., bias). Approximate sampling distributions are given for the test statistic. Two modes of use are distinguished, an exploratory mode in which items are identified for further scrutiny, and a confirmatory mode in which the method may be applied to individual items with higher precision. Demonstrations with actual and simulated data are reported.

AFHRL Technical Paper 83-33

September 1983

IDENTIFYING DIFFERENT ITEM RESPONSE CURVES

By

Michael V. Levine

University of Illinois
Urbana-Champaign, Illinois 61820

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235

Reviewed and submitted for publication by

Janos B. Koplyay
Chief, Manpower and Force Management Systems Branch
Manpower and Personnel Division

This publication is primarily a working paper.
It is published solely to document work performed.

Acknowledgments

This project was supported by contract number F41698-81-C-0012. The Principal Investigator for McFann-Gray & Associates, Inc. was Benjamin Fairbank, Jr. Appreciation for recognizing the need for and supporting this research is due to Dr. Janos Koplyay and Dr. Malcolm James Ree. The research reported develops, extends, and implements theoretical and numerical results supported by the Office of Naval Research contract N00014-79C-0752; NR-150445, to be reported separately. Valuable conversations with Benjamin Fairbank, C. Nicholas Hastings, Robert Linn, Daniel Segall, James Wardrop and Bruce Williams on this work are gratefully acknowledged.

Table of Contents

List of Tables

Abstract

A method is presented for determining whether or not two three-parameter item characteristic curves differ significantly from each other. The method may be used within the context of item response theory to detect evidence of item compromise, change with time, or group-specific differences (e.g., bias). Approximate sampling distributions are given for the test statistic. Two modes of use are distinguished, an exploratory mode in which items are identified for further scrutiny, and a confirmatory mode in which the method may be applied to individual items with higher precision. Demonstrations with actual and simulated data are reported.

The method involves the derivation of a measure analogous to chi-square, called a sum of squares (SOS) measure. The SOS measures integrate the difference between item response curves and assess the statistical significance of the resulting area based not only on its magnitude but also on the accuracy with which the two curves were estimated.

Identifying Different Item Response Curves

Introduction

Independent samples are drawn from two populations and administered the same test. Item response curves are estimated for each sample and compared. For a given item, the graphs of the estimated curves may look quite different across samples. When can one safely conclude that the population item response functions differ? In other words, when can one infer from estimated curves that the same item functions differently in the two populations?

This issue, which originates in item bias studies, arises naturally in many other settings. If security has failed in one testing center for one or a small number of items, then large differences in estimated item response functions for the compromised items are expected between centers. If the format, wording, or position of an item in a test is changed, then did the item change functionally? If an item is moved from one test to a new test, is it still functionally the same item? If an item is restored to use after a decade, have changes in language, educational practice, or society made the item harder or less discriminating? Has a revision of a parameter estimation program resulted in reliable differences in estimated curves?

The research to be reported is an advance over earlier attempts to compare item characteristic curves. Sampling distributions of measures of differences between curves are approximated. A technique is introduced to permit the user to focus on portions of curves of special interest.

A class of statistics for comparing curves, Sum of Squares (SOS) statistics, is identified and analyzed. Exploratory and confirmatory research designs for comparing curves are distinguished. Finally, empirical studies with actual and with simulated data illustrating and validating the theoretical results are reported.

SOS Statistics

The discrepancy between two curves f and g can be quantified by summing or integrating squared differences between curves. For example summing over, say, 600 closely spaced points gives the index

$$\sum_{n=1}^{600} [f(-3+0.01n) - g(-3+0.01n)]^2$$

that was used in Linn, Levine, Hastings, and Wardrop, 1980. Essentially the same theory obtains for the integrals approximated by such sums, such as

$$100\int_{-3}^{3} [f(t) - g(t)]^2 dt$$

and more generally

$$\int_{-\infty}^{\infty} [f(t) - g(t)]^2 \, w(t)\, dt$$

where w(t) can be specified as

$$\begin{aligned} w(t) &= 100, \text{ for } -3 \leq t \leq 3 \\ &= 0, \text{ otherwise.} \end{aligned}$$

The results in this paper are also applicable to more complicated "weight" functions w(t).

Measures of this type will be called "Sum of Squares" indices or, more briefly, SOS measures. The non-negative weight function w controls the contribution of portions of the curve to the measure. When $\hat{f}$ and $\hat{g}$ are estimated curves, then the summation formula

$$\sum [\hat{f}(\theta_i) - \hat{g}(\theta_i)]^2 \, w(\theta_i)$$

and the integral formula

$$\int [\hat{f}(\theta) - \hat{g}(\theta)]^2 \, w(\theta)\, d\theta$$

define statistics. Such statistics will be called SOS statistics below.

Having the option to choose weights is important. Some portions of estimated curves are more important tha╕ others and some portions, better estimated. If a test is being used to select enlistees for advanced technical training it is important to have good measurement over high ability ranges. If the calibrating sample contains only high ability examinees and no low ability examinees then the lower portions of the curves will depend on extrapolation only.

In this report constant weights were used only to obtain two SOS formulas

$$\text{(a)} \int_{-3}^{3} [\, f(\theta) - g(\theta)]^2 \, d\theta$$

$$\text{(b)} \int_{-2.5}^{1} [\, f(\theta) - g(\theta)]^2 \, d\theta$$

Formula (a) has been used extensively, especially by Linn, Levine, Hastings, and Wardrop (1981), whose data are reanalyzed. Formula (b) seems more appropriate for the portions of their data reanalyzed, because very few high ability examinees are included in the data and because the test has a low information function (Lord, 1980) over the low ability range. If the interval of θ is from -3 to +3, the range will include virtually all cases found in the ability distribution whose mean is 0 and standard deviation is 1. A range of -2.5 to 1 includes virtually all cases when the distribution is skewed, as in the studies reported below. It omits poorly measured low ability cases.

Background

In this section some of the statistics and mathematics supporting this report are reviewed. Let $P(\theta ; \underline{\alpha})$ be the probability that an examinee sampled from all those with ability θ correctly answers an item with parameters $\underline{\alpha}$ (underlining indicates that $\underline{\alpha}$ is a vector) θ is a scaler and $\underline{\alpha}$ is a (row) vector of item parameters $<a, b, c>$, P is a three-parameter logistic function, defined as

$$P(\theta ; \alpha) = c + (1 - c) / [\, 1 - e^{-a(\theta - b)}] . \tag{1.1}$$

In a typical study $\underline{\alpha}$ will be estimated from separate samples to yield two estimates $\hat{\underline{\alpha}}_1$ and $\hat{\underline{\alpha}}_2$ of the correct value of $\underline{\alpha}$, $\underline{\alpha}_0$; This section develops an approximation for the integral form of the general SOS statistic, defined as

$$\int_{-\infty}^{\infty} [\, P(\theta ; \hat{\underline{\alpha}}_1) - P(\theta ; \hat{\underline{\alpha}}_2) \,]^2 \, w(\theta) \, d\theta . \tag{1.2}$$

The weight function w is non-negative and non-zero only over an interval. (Note: the notation $\underline{\alpha}_1$, and $\underline{\alpha}_2$ indicate, as before, the first and second alpha vectors, not the first and second elements of the vector, alpha. The underlining continues to denote the vector.) When P has continuous partial derivatives the mean value theorem can be used to develop a tractable approximation of (1.2). In the three-parameter logistic model

(1.3)
$$P(\theta;\hat{\underline{\alpha}}_1) - P(\theta;\hat{\underline{\alpha}}_2) =$$
$$(\hat{a}_1 - \hat{a}_2)\frac{\partial}{\partial a} P(\theta;\underline{\alpha}^*)$$
$$+ (\hat{b}_1 - \hat{b}_2)\frac{\partial}{\partial b} P(\theta;\underline{\alpha}^*)$$
$$+ (\hat{c}_1 - \hat{c}_2)\frac{\partial}{\partial c} P(\theta;\underline{\alpha}^*)$$

Here $\underline{\alpha}^*$ is a vector on the line segment connecting $\hat{\underline{\alpha}}_1$ and $\hat{\underline{\alpha}}_2$,

(1.4)
$$\underline{\alpha}^* = h\hat{\underline{\alpha}}_1 + (1-h)\hat{\underline{\alpha}}_2 \qquad 0 < h < 1.$$

The value of h needed in (1.4) will depend on θ . In this study it has been observed that P is so close to being linear over the range between $\hat{\underline{\alpha}}_1$ and $\hat{\underline{\alpha}}_2$ that an adequate approximation can be obtained for h = 0.5 for all values of θ .

Equation (1.3) can be expressed more compactly with the notation

$$P(\theta;\hat{\underline{\alpha}}_1) - P(\theta;\hat{\underline{\alpha}}_2) = (\hat{\underline{\alpha}}_1 - \hat{\underline{\alpha}}_2)\frac{\partial}{\partial\underline{\alpha}} P(\theta;\underline{\alpha}^*(\theta))$$

where $\frac{\partial}{\partial\underline{\alpha}} P(\theta;\underline{\alpha}^*(\theta))$ is a column vector of partial derivatives evaluated at

$h(\theta)\hat{\underline{\alpha}}_1 + [1 - h(\theta)]\hat{\underline{\alpha}}_2$. Thus equation (1.3) in this notation becomes

$$P(\theta;\hat{\underline{\alpha}}_1) - P(\theta;\hat{\underline{\alpha}}_2) = (\hat{\underline{\alpha}}_1 - \hat{\underline{\alpha}}_2)\frac{\partial}{\partial\underline{\alpha}} P(\theta;.5\hat{\underline{\alpha}}_1 + .5\hat{\underline{\alpha}}_2)$$

$$= (\hat{\underline{\alpha}}_1 - \hat{\underline{\alpha}}_2)\frac{\partial}{\partial\underline{\alpha}} P(\theta;\bar{\hat{\underline{\alpha}}})$$

where $\bar{\hat{\underline{\alpha}}} = 0.5\hat{\underline{\alpha}}_1 + 0.5\hat{\underline{\alpha}}_2$ is the mean $\underline{\alpha}$ estimate.

With this approximation an SOS statistic can be written compactly as

$$(\hat{\underline{\alpha}}_1 - \hat{\underline{\alpha}}_2)Q(\hat{\underline{\alpha}}_1 - \hat{\underline{\alpha}}_2)^T.$$

Here superscript T indicates vector (and later, matrix) transposition and Q is the positive semidefinite matrix obtained by integrating with respect to θ each term in the matrix

$$\left(\frac{\partial}{\partial\underline{\alpha}} \bar{P}(\theta ; \hat{\underline{\alpha}})\right)\left(\frac{\partial}{\partial\underline{\alpha}} \bar{P}(\theta ; \hat{\underline{\alpha}})\right)^T w(\theta)$$

As evidence for the adequacy of this approximation Table 1 is offered. In this table 45 items are considered. Item parameters were estimated for Black and for white examinees. The SOS statistic

$$\int_{-3}^{3} [P(\theta; \hat{\underline{\alpha}}_1) - P(\theta ; \hat{\underline{\alpha}}_2)]^2 d\theta$$

was computed numerically (Riemann sums on a fine grid) and compared with the quadratic approximation. These results are typical. Comparing columns two and three in Table 1, respectively the integral and quadratic forms, close agreement between the integral and its quadratic approximation was observed. (The eigenvalues in the table will be referred to later.) Details of the data set may be found in Linn et al., 1980.

When the estimates $\hat{\underline{\alpha}}_1$ and $\hat{\underline{\alpha}}_2$ are multivariate normal with known covariance matrices S_1 and S_2 , then the random difference vector $(\hat{\underline{\alpha}}_1 - \hat{\underline{\alpha}}_2)$ will be multivariate normal with covariance matrix $S = S_1 + S_2$, provided the estimates $\hat{\underline{\alpha}}_1$ and $\hat{\underline{\alpha}}_2$ are obtained from independent samples.

If $\hat{\underline{\alpha}}_1$ and $\hat{\underline{\alpha}}_2$ are unbiased estimates of the correct item parameters $\underline{\alpha}_0$, (where $\underline{\alpha}_0$ indicates the true (unknown) population values, which $\underline{\alpha}_1$ and $\underline{\alpha}_2$ estimate) then the expected value of the difference $\hat{\underline{\alpha}}_1 - \hat{\underline{\alpha}}_2$ will be zero.

An SOS statistic has approximately the same distribution as the sum of several independent squared normal variables. The present line of reasoning is intended

Table 1.

Quadratic approximation of integral
for 45 items. Parameters were computed from white fifth grade and Black sixth grade samples (Linn, Levine, Hastings, & Wardrop, 1980).

Item	Integral	Quadratic Form	Eigenvalues of $D^{\frac{1}{2}}V^TQVD^{\frac{1}{2}}$		
1	0.00749	0.00750	0.00318	0.00109	0.00032
2	0.01269	0.01280	0.00332	0.00132	0.00032
3	0.06779	0.06913	0.00809	0.00544	0.00116
4	0.00555	0.00559	0.00361	0.00140	0.00103
5	0.01806	0.01823	0.00274	0.00156	0.00098
6	0.00433	0.00429	0.00303	0.00119	0.00035
7	0.00051	0.00051	0.00292	0.00116	0.00025
8	0.01985	0.02005	0.00234	0.00086	0.00043
9	0.03331	0.03432	0.00608	0.00259	0.00129
10	0.00348	0.00348	0.00258	0.00095	0.00042
11	0.00295	0.00293	0.00230	0.00084	0.00049
12	0.01015	0.01021	0.00560	0.00140	0.00092
13	0.00303	0.00304	0.00258	0.00168	0.00087
14	0.01319	0.01329	0.00214	0.00098	0.00029
15	0.00391	0.00391	0.00277	0.00114	0.00025
16	0.00120	0.00119	0.00276	0.00106	0.00034
17	0.03035	0.02718	0.00279	0.00221	0.00107
18	0.00538	0.00538	0.00186	0.00109	0.00085
19	0.00722	0.00725	0.00151	0.00079	0.00028
20	0.01250	0.01254	0.00167	0.00077	0.00053
21	0.00570	0.00570	0.00228	0.00087	0.00048
22	0.02531	0.02306	0.00240	0.00179	0.00097
23	0.00564	0.00557	0.00240	0.00087	0.00064
24	0.00203	0.00203	0.00213	0.00106	0.00083
25	0.02476	0.02464	0.06258	0.00329	0.00254
26	0.03518	0.03568	0.00153	0.00075	0.00067
27	0.00751	0.00757	0.00469	0.00145	0.00101
28	0.00833	0.00837	0.00258	0.00098	0.00035

Table 1, continued.

29	0.00240	0.00237	0.00165	0.00080	0.00068
30	0.01085	0.01089	0.00363	0.00146	0.00121
31	0.02497	0.02577	0.00795	0.00178	0.00142
32	0.00429	0.00428	0.00177	0.00141	0.00084
33	0.00383	0.00376	0.00164	0.00089	0.00066
34	0.03503	0.03550	0.00273	0.00107	0.00086
35	0.01222	0.01228	0.01083	0.00234	0.00104
36	0.03942	0.03769	0.00699	0.00252	0.00207
37	0.00484	0.00485	0.00139	0.00105	0.00077
38	0.00567	0.00568	0.00178	0.00149	0.00097
39	0.00737	0.00744	0.00165	0.00097	0.00082
40	0.03315	0.03398	0.01101	0.00248	0.00121
41	0.00605	0.00607	0.00632	0.00194	0.00119
42	0.00433	0.00435	0.01019	0.00226	0.00133
43	0.04684	0.04829	0.03279	0.00614	0.00131
44	0.00450	0.00452	0.00251	0.00114	0.00082
45	0.00918	0.00934	0.00191	0.00154	0.00104

to make this more specific. Let $\underline{\alpha}$ temporarily denote an arbitrary multivariate normal random vector, and let Q temporarily denote a given (not estimated) positive semidefinite (i.e., non-negative latent roots) symmetric matrix. If the following conditions hold for $\underline{\alpha}$ and some matrices,

$$E\,(\underline{\alpha}) = 0$$

$$\mathrm{Cov}\,(\underline{\alpha}) = E\,(\underline{\alpha}^T \underline{\alpha}) = S$$

$S = VDV^T$ where V is orthonormal and D is diagonal with positive diagonal elements

$$D^{\frac{1}{2}} = \text{matrix of square roots of elements of D}$$

$$D^{-\frac{1}{2}} = (D^{\frac{1}{2}})^{-1}$$

then the transformed random vector $\underline{\beta} = \underline{\alpha} V D^{-\frac{1}{2}}$ is a vector of independent standard normal variables. This follows from the identities

$$\begin{aligned} E\,(\underline{\beta}) &= E\,(\underline{\alpha})\, VD^{-\frac{1}{2}} = 0 \\ \mathrm{Cov}\,(\underline{\beta}) &= E\,[D^{-\frac{1}{2}} V^T \underline{\alpha}^T \underline{\alpha} VD^{-\frac{1}{2}}] \\ &= D^{\frac{1}{2}} V^T E\,(\underline{\alpha}^T \underline{\alpha})\, VD^{-\frac{1}{2}} \\ &= D^{-\frac{1}{2}} V^T S V D^{-\frac{1}{2}} \\ &= I\,. \end{aligned}$$

A statistic $\underline{\alpha}\, Q \underline{\alpha}^T$ thus can be rewritten

$$\begin{aligned} \underline{\alpha}\, Q \underline{\alpha}^T &= \underline{\alpha}(VD^{-\frac{1}{2}} D^{\frac{1}{2}} V^T)\, Q\, (\underline{\alpha} VD^{-\frac{1}{2}} D^{\frac{1}{2}} V^T) \\ &= (\underline{\alpha}\, VD^{-\frac{1}{2}})\, D^{\frac{1}{2}} V^T Q V D^{\frac{1}{2}} (\underline{\alpha}\, VD^{\frac{1}{2}})^T \\ &= \underline{\beta}\, D^{\frac{1}{2}} V^T Q V D^{\frac{1}{2}} \underline{\beta}^T \end{aligned}$$

where $\underline{\beta}$ is a vector of independent standard normal variables. If U diagonalizes $D^{\frac{1}{2}}V^TQVD^{\frac{1}{2}}$, i.e., if U is an orthonormal matrix and

$$\Lambda = \begin{pmatrix} \lambda_1 & o & o \\ o & \lambda_2 & o \\ o & o & \lambda_3 \end{pmatrix}$$

is a diagonal matrix such that

$$D^{\frac{1}{2}}V^TQVD^{\frac{1}{2}} = U \Lambda U^T,$$

then $\underline{\alpha} Q \underline{\alpha}^T = (\underline{\beta} U) \Lambda (\underline{\beta} U)^T$. $\underline{\beta} U = < x_1 , x_2 , x_3 >$, being an orthonormal transformation of independent standard normal variables, is also a vector of independent standard normal variables. This establishes that $\underline{\alpha} Q \underline{\alpha}^T$ has the same distribution as the random variable $\lambda_1 x_1^2 + \lambda_2 x_2^2 + \lambda_3 x_3^2$ where the x_i are independent standard normal variables and the λ_i are eigenvalues of $D^{\frac{1}{2}}V^TQVD^{\frac{1}{2}}$.

The above is used in the following form. If $\hat{\underline{\alpha}}_1$ and $\hat{\underline{\alpha}}_2$ are independent multivariate normal vectors with covariance matrices S_1 and S_2 and equal expected values, and if Q is symmetric positive semidefinite, then $\hat{\underline{\alpha}}_1 - \hat{\underline{\alpha}}_2$ is multivariate normal with zero expectation and covariance matrix S equal to $S_1 + S_2$. The random variable $(\hat{\underline{\alpha}}_1 - \hat{\underline{\alpha}}_2)\ Q\ (\hat{\underline{\alpha}}_1 - \hat{\underline{\alpha}}_2)^T$ will have the same distribution as the variable $\sum \lambda_i x_i^2$ where the x_i are independent standard normal variables and the λ_i, in the notation of the preceding paragraph, the eigenvalues of $D^{\frac{1}{2}}V^TQVD^{\frac{1}{2}}$.

This result is important because it shows that an SOS statistic has approximately the same distribution as a homogeneous quadratic form in normal variables. These variables generalize the central chi square distribution. There are no tables for the variables, but there is literature (Johnson and Kotz, 1967) which provides a variety of methods for computing their distribution.

In this report the estimates of item parameters $\hat{\underline{\alpha}}_1, \hat{\underline{\alpha}}_2$ are approximations of maximum likelihood estimates for an item's parameters (a, b, and c). The methods here reported are also valid for other estimates (such as Bayesian) if the estimates show asymptotic normality. This validity holds because the derivation does not use maximum likelihood, but only multivariate normality. The estimates $\hat{\underline{\alpha}}_1$ and $\hat{\underline{\alpha}}_2$ are based on samples from different populations. The covariance matrices S_1 and S_2 are obtained by inverting information matrices. The condition $E(\hat{\underline{\alpha}}_1) = E(\hat{\underline{\alpha}}_2)$ is obtained from the "null hypothesis" of no bias, i.e., that the correct conditional probability curve is the same for each population, and the matrix Q is obtained by the mean value theorem approximation. To calculate the probability of obtaining an observed statistic under the null hypothesis, a recently developed numerical procedure for inverting the Laplace transformation of the distribution of the statistic is used (Levine and Williams, 1982).

Every one of the above conditions can be questioned; every one of the assumptions and approximations can be refined. For example, the approximate maximum likelihood estimates $\hat{\underline{\alpha}}_1$, $\hat{\underline{\alpha}}_2$ are statistically biased and the matrix Q is estimated from data. Actual and simulation data are therefore needed for an evaluation of the extent to which the assumptions and approximations result in useful methods.

Exploratory vs. Confirmatory Studies

Two anticipated uses of SOS statistics are exploratory and confirmatory. In some situations a low power, easily implemented exploratory index is needed to screen for items requiring further investigation. Other situations call for the use of a precise test to confirm or reject a hypothesis about one or a small number of items.

Consider the problem of item security. An inexpensive exploratory test makes possible routine, periodic screening for compromised items. If an item has been disclosed, its item response function will be different when a sample is drawn from a secure test center or from a population tested prior to disclosure. After an exploratory study has tagged a particular item as possibly compromised, the item can be grouped with safe items for the more precise confirmatory study.

The steps in an exploratory study are:

(E1) Estimate item response functions from two samples drawn from two populations.

(E2) Estimate covariance matrices for the estimated item parameters.

(E3) Equate the two populations and place the curves on a common ability scale.

(E4) Compute the SOS statistic for each pair of estimated curves.

(E5) Calculate the probability of observing a value of SOS as large or larger than the sample SOS under the null hypothesis that the two curves are equal.

Steps (E1) - (E4) are essentially the same as those in Linn et al., (1980, 1981). Step (E5) is an attempt to measure the significance and replicability of large SOS values.

The equating step (E3) is especially problematical in exploratory studies, when ability distributions are markedly different. Confirmatory studies circumvent equating problems. The principal steps in confirmatory studies are described below for the important special case in which only one item is suspected of being compromised.

(C1) Merge item response data from two independent samples for a subtest consisting of unbiased items.

(C2) Estimate abilities from the merged file.

(C3) Calculate maximum likelihood item parameter estimates, treating estimated abilities as actual abilities separately for each population on the suspected item.

(C4) Compute the covariance matrix for the item parameter estimates.

(C5) Same as (E4)

(C6) Same as (E5)

Note that in the confirmatory study estimation using an unbiased subtest automatically equates and places estimated abilities on a common scale. In fact, if the merged sample of examinees has a much broader range of ability than either component samples, then the quality of subtest item parameter estimates is likely to be improved, and the ability estimates will be more precise. In this sense, in the confirmatory study, ability distribution differences can be an asset rather than a liability. Furthermore, in the confirmatory study there is no "multiple comparison problem," and the "large sample theory" used to deduce asymptotic multivariate normality of parameter estimates can be shown to be more nearly correct.

Further details on exploratory and confirmatory methods are given in the empirical sections of this report.

Empirical Studies: Introduction and Overview

In the remainder of this report a sequence of empirical studies with reading test data is described. These studies were undertaken to make an initial determination of whether or not the procedure developed here would function as the derivations indicated when exposed to operational data. The first, an exploratory study with actual data, is a partial replication of Linn et al.(1981). Three items were identified as possibly biased in that study. The second, an exploratory study with simulated data having roughly the same ability distribution and item parameters as the first study, demonstrates that isolated biased items can be separated from unbiased items by SOS statistics. In the study, an unbiased item, item 7, was incorrectly identified as biased. In the third study, a confirmatory study with the same simulation parameters as the second study, the three biased items (numbers 6, 19, and 26) were clearly identified as biased and the incorrectly identified item (7) was correctly reclassified as unbiased. In a final confirmatory study with actual data, a strong indication of bias was obtained for one of the three original items.

This report deals only with populations defined by educational achievement. If performance on an item requires two component skills, one of which is commonly taught during the fifth grade, then, under a variety of complex but obvious conditions, estimated item characteristic curves should reliably differ for fifth and sixth grade students. It is preferred to strengthen the understanding of these new methods in this uncontroversial domain before presenting results on populations defined by race or income level.

Exploratory Study: Actual Data

The first study replicates Linn et al.(1980, 1981) with minor changes in their methods and a sample from the same data set. A spaced sample of 1940 examinees was selected from the 2910 available low income, white, fifth grade (LW5) examinees completing Form F (45 items) of the Metropolitan Achievement

Test (Durost, Bixler, Wrightstone, Prescott, & Balow, 1970), Reading Comprehension section. A second sample of 1948 (selected from 2752) low income, white, sixth grade (LW6) examinees was also formed. Item parameters were computed for each sample using LOGIST (Wood, Wingersky, and Lord, 1976). Item parameter covariance matrices were approximated by inverting matrices of sample averages of second partial derivatives (Linn et al. 1980, Appendix A). Four examinees from the LW6 sample with such unusual test performance that LOGIST failed to estimate an ability in the interior of the interval from $\theta = -4$ to $\theta = +4$ were excluded from the averages, leaving a sample of 1944. The four examinees on whom convergence failed were not used in any subsequent steps. The LW6 abilities and item parameters were linearly transformed to place them on approximately the same scale as the LW5 parameters. (See Linn et al. (1980, Appendix B) for details of the equating procedure.)

Item characteristic curves were compared with the SOS statistic obtained by integrating the squared difference between unit weighted curves where $-3 \leq \theta \leq +3$. Adequate agreement with the earlier results (Linn et al. 1980) was observed. A separate report of the replication is anticipated.

After examining the cumulative distribution function of the estimated abilities it was decided to consider integrating the difference between curves between -2.5 and 1 only, where most cases were concentrated. Only 5 percent of the LW6 sample obtained scores greater than 1. The major SOS statistic for this sequence of studies was

$$\int_{-2.5}^{1} [P(\theta\, ; \underline{\hat{\alpha}}_1) - P(\theta\, ; \underline{\hat{\alpha}}_2)]^2 \, d\theta\, .$$

Table 2 gives the SOS values and probabilities for selected items. Items 6, 19, and 26 showed the largest SOS values. The probability of observing the obtained or larger SOS values in each case was estimated to be less than 0.01. These were the only items with significance less than 0.01, and they were selected for special attention in the sequel. The item with the next largest SOS value was item 31 with p > .10, which is not considered statistically significant. Table 3 gives item parameters for LW5 and equated LW6 parameters. Note that 28 of the 45 LW5 values for the c parameter (those where c = 0.235) and 33 of the 45

Table 2
Exploratory LW5 vs. LW6, actual data.
SOS values and their probabilities.

Item	SOS Statistic	Probability (P)
1	0.00489	0.728
2	0.00246	0.439
3	0.00030	0.028
4	0.00008	0.007
5	0.00477	0.701
6	0.03361	0.999
7	0.00745	0.811
8	0.00224	0.353
9	0.00224	0.380
10	0.00492	0.679
11	0.00653	0.783
12	0.00669	0.879
13	0.00092	0.136
14	0.00444	0.637
15	0.00634	0.756
16	0.00495	0.656
17	0.00924	0.917
18	0.00748	0.971
19	0.01881	0.995
20	0.00866	0.916
21	0.00042	0.041
22	0.00048	0.065
23	0.00048	0.051
24	0.00153	0.281
25	0.00382	0.493
26	0.02077	0.998
27	0.00179	0.497
28	0.00567	0.748
29	0.00021	0.017

Table 2, continued.

30	0.00142	0.407
31	0.01311	0.895
32	0.00217	0.496
33	0.00259	0.516
34	0.00133	0.398
35	0.00084	0.371
36	0.00086	0.193
37	0.00084	0.158
38	0.00260	0.578
39	0.00759	0.895
40	0.00035	0.116
41	0.00113	0.387
42	0.00147	0.255
43	0.00055	0.237
44	0.00473	0.669
45	0.00155	0.314

Table 3

Estimated item parameters from LW5/LW6.

Exploratory study on the LW5 ability scale.

	LW5			LW6		
Item	a_i	b_i	c_i	a_i	b_i	c_i
1	0.675	-1.621	0.235	0.747	-1.413	0.230
2	0.757	-1.436	0.235	0.728	-1.611	0.230
3	0.426	0.652	0.235	0.377	0.720	0.230
4	1.364	1.044	0.235	1.295	1.031	0.230
5	0.710	-0.203	0.235	0.659	0.005	0.230
6	0.868	-0.882	0.235	0.786	-1.435	0.230
7	1.200	-1.101	0.235	1.016	-1.305	0.230
8	1.299	-0.288	0.235	1.068	-0.347	0.230
9	0.540	0.316	0.235	0.652	0.148	0.230
10	0.966	-0.439	0.235	0.912	-0.631	0.230
11	1.363	-0.133	0.260	1.028	-0.339	0.230
12	1.293	1.452	0.243	0.806	1.387	0.230
13	0.679	0.230	0.235	0.616	0.334	0.230
14	1.574	-0.811	0.142	1.921	-0.721	0.230
15	1.563	-0.969	0.235	1.295	-0.840	0.230
16	1.032	-0.823	0.235	0.848	-0.977	0.230
17	0.822	0.729	0.220	0.638	0.608	0.230
18	2.000	1.000	0.186	1.796	1.004	0.230
19	2.000	-0.374	0.171	1.947	-0.511	0.230
20	1.956	0.285	0.303	1.309	0.108	0.230
21	1.364	-0.246	0.235	1.308	-0.216	0.230
22	0.901	0.709	0.235	0.829	0.661	0.230
23	0.988	0.012	0.235	1.060	-0.025	0.230
24	1.037	0.292	0.220	1.160	0.382	0.230
25	0.515	2.119	0.235	1.191	1.561	0.255
26	1.303	0.313	0.145	1.073	0.250	0.210
27	1.871	1.185	0.281	1.947	1.095	0.297
28	1.172	-0.464	0.235	1.399	-0.348	0.230

Table 3, continued.

29	1.570	0.372	0.235	1.619	0.343	0.230
30	1.401	1.285	0.239	1.850	1.125	0.260
31	0.879	1.556	0.235	0.458	1.743	0.230
32	1.519	0.756	0.235	1.541	0.871	0.260
33	1.903	0.378	0.235	1.593	0.420	0.210
34	2.000	0.883	0.160	1.723	0.957	0.146
35	2.000	1.738	0.235	1.944	1.476	0.230
36	0.915	1.485	0.235	0.905	1.328	0.230
37	2.000	0.554	0.220	1.947	0.588	0.210
38	2.000	0.878	0.287	1.886	0.913	0.260
39	1.841	0.392	0.235	1.446	0.326	0.260
40	2.000	1.291	0.220	1.670	1.408	0.210
41	2.000	1.492	0.235	1.690	1.518	0.247
42	0.750	1.214	0.235	0.616	1.210	0.230
43	2.000	1.622	0.243	1.947	1.592	0.230
44	1.215	-0.084	0.235	1.373	0.033	0.230
45	1.876	0.829	0.235	1.390	0.833	0.230

LW6 values for the c parameter (those where $c = 0.23$) are assigned by LOGIST to approximated default values. Such assignments affect asymptotic normality and so are part of the set of violations of the assumptions (page 11) whose effect the simulation is evaluating.

Exploratory Study: Simulation Data

The purpose of this study was to determine whether SOS statistics can identify isolated biased items under ideal conditions. Data files SLW5 and SLW6 were constructed by simulation to parallel LW5 and LW6. The abilities used for SLW5 were the abilities estimated by LOGIST from LW5. The abilities used for SLW6 were the abilities estimated for LW6 transformed to the LW5 scale by the empirical equating transformation derived in the earlier study $\theta \rightarrow (1.027)\,\theta + 0.399$. This is a linear transformation which does not change the meaningfulness of the ability measurements, since measurement based on item response theory is unique only up to a linear transformation.

The 1999 abilities were obtained for the SLW5 simulation by using all 1940 estimated θ's from the LW5 at least once and the first 59 θ 's twice. The 1999 abilities for the SLW6 simulation were obtained by using all 1944 transformed estimated LW6 θ 's at least once and the first 55 twice.

Except as indicated below, the item parameters for both SLW5 and SLW6 were the LOGIST estimated LW6 item parameters transformed to the LW5 ability scale by the following transformations, which are derived from the transformation given above for the theta values.

$$a_i \rightarrow a_i \quad (1.027)$$

$$b_i \rightarrow (1.027)\, b_i + 0.399.$$

$$c_i \rightarrow c_i$$

Except for items 6, 19, and 26 all c_i's were set equal to 0.21, a value typical of those found for the c parameter when convergence succeeds with large samples. The nine parameters for items 6, 19, and 26 for SLW5 were the estimated LW5 parameters. The nine parameters for items 6, 19, and 26 for SLW6 were the transformed estimated LW6 parameters. Only a_i's less than or equal to 1.7 were used. Any a_i found to be greater than 1.7 by the above procedure was changed to 1.7.

Table 4 gives the SOS statistic values and their probabilities. All three of the critical items were clearly identified as biased at the 0.01 level. The SOS values were 0.036 (item 6), 0.035 (item 19), 0.023 (item 26). Only one of the unbiased items, item 7, with the very large SOS value of 0.043 was significant at the 0.01 level. The remaining items had generally small SOS values, the largest of which was only 0.006.

Confirmatory Study: Simulation Data

The purpose of a confirmatory study is to confirm or disconfirm the suspicion of bias raised for items in an exploratory study. A merged file of 1998 simulees was formed of 999 simulated LW5 examinees and 999 simulated LW6 examinees. The abilities were the 999 thetas in each file, LW5 and LW6, which followed serially the last of the abilities used for the exploratory simulations (page 19).

In constructing the merged file a process called "cloning" was used. A 49 item, rather than a 45 item, test was constructed as follows. All LW5 simulees were coded as not having reached items 46, 47, 48, and 49. The LW6 examinees were coded as not having reached items 6, 7, 19, and 26. The response to item 6 was moved to the 46th position, item 7 to the 47th position, etc. In this way the abilities are held to a common scale by the 41 common items, but LOGIST is free to fit different parameters for each group for the cloned items.

This procedure was used in order to evaluate the confirmatory design with readily available software. It would have been preferable, although infeasible, to run LOGIST on the 41 common items and fit each of the special items separately. The procedure actually used permits the possibly biased items to influence the ability estimates.

In order to make the probabilities in this study comparable to the probabilities in the exploratory study, a correction for sample size was used. The covariance matrix of the estimated parameters (under ideal conditions, such as perfectly estimated abilities) are inversely proportional to sample size. The

Table 4

Exploratory study, simulation data, SOS value and probabilities. (Rounding has produced numbers of 1 and 0. The + and - signs indicate that such values have been rounded and are not accurate to the last decimal.)

Item	SOS Statistic	Probability (P)
1	0.00582	0.803
2	0.00202	0.394
3	0.00081	0.125
4	0.00095	0.272
5	0.00015	0.009
6	0.03641	0.999*
7	0.04337	1.000-*
8	0.00137	0.213
9	0.00515	0.702
10	0.00010	0.004
11	0.00015	0.008
12	0.00245	0.548
13	0.00040	0.043
14	0.00027	0.020
15	0.00046	0.043
16	0.00084	0.131
17	0.00096	0.167
18	0.00127	0.365
19	0.03465	1.000-*
20	0.00004	0.001
21	0.00188	0.338
22	0.00029	0.033
23	0.00326	0.549
24	0.00433	0.720
25	0.00068	0.208
26	0.02268	0.999*
27	0.00529	0.922
28	0.00193	0.328
29	0.00370	0.709

Table 4, continued.

30	0.00325	0.776
31	0.00025	0.033
32	0.00091	0.229
33	0.00246	0.523
34	0.00294	0.711
35	0.00120	0.467
36	0.00002	0.001
37	0.00151	0.354
38	0.00298	0.732
39	0.00330	0.640
40	0.00000+	0.000+
41	0.00150	0.557
42	0.00004	0.002
43	0.00034	0.166
44	0.00232	0.438
45	0.00384	0.773

* = $p \leq 0.01$

exploratory studies have sample sizes equal to 1999. Therefore the estimated covariance matrices were multiplied by actual size and divided by 1999 to compensate for sample size.

The results confirm bias at the 0.01 level for the truly biased items. The "false alarm item" (number 7) no longer appears biased, and its SOS value has a probability of occurring by chance of over 0.13. The results are summarized in Table 5. Note that the probability associated with item 7 is .8683, or approximately 1-0.13. Thus the test appears to have high power at normally acceptable alpha levels.

Confirmatory Study: Actual Data

All of the remaining LW5 and LW6 examinees were merged for a confirmatory study. This produced sample sizes somewhat smaller than the sample sizes in the simulation confirmatory studies: 970 LW5 examinees and 804 LW6 examinees. Items 6, 19, and 26 were cloned, and the procedures described in the preceding section, implemented. Results with sample sizes corrected to 999 and 1999 are presented in Table 6.

The evidence for bias is strongest for item 26. The estimated item parameters $<a, b, c>$ for the fifth and sixth graders were $<1.06, -0.032, 0.09>$ and $<1.18, 0.063, 0.21>$. The discrepancy between the estimated curves seems attributable primarily to the discrepant estimates of c_i. (See Table 7 for item response functions and their confidence intervals.) It is tempting to conjecture that one of the distractor options ceases to be effective after fifth grade instruction.

Conclusions

One mode of the procedure described in this report may be used to develop SOS statistics which identify items which appear to be biased. That mode is the exploratory mode. The other mode, the confirmatory mode, can be used to examine the suspect items and to determine, with relatively high statistical power, whether the hypothesis of no difference between the items can be rejected at specified alpha levels. The probabilities of observed SOS values

Table 5

SOS values and their associated probability values for items suspected of being biased, simulated data.

Item	SOS	Probability corrected for sample size	Probability uncorrected for sample size
6	0.08507	0.9999998	0.9998
7	0.00764	0.8683	0.6453
19	0.01302	0.9907	0.9104
26	0.17929	0.9977	0.9615

Table 6

SOS values and their associated probability values
for items suspected of being biased, actual data.

Item	SOS	Probability corrected to sample size of: 1999	999
6	0.0123	0.944	0.789
19	0.0013	0.209	0.095
26	0.0166	0.997	0.957

Table 7

Conditional probabilities $P(1|\theta)$ and their confidence intervals for item 26, confirmatory study, actual data, LW5 vs. LW6.

	θ:	-3.0	-2.0	-1.0	0	1	2.0	3.0
	$P(1|(\theta - 2\sigma_\theta))$	0.02	0.06	0.18	0.51	0.84	0.96	0.99
LW5	$P(1|\theta)$	0.10	0.12	0.23	0.56	0.88	0.98	1.00
	$P(1|(\theta + 2\sigma_\theta))$	0.17	0.17	0.27	0.60	0.92	0.99	1.00
	$P(1|(\theta - 2\sigma_\theta))$	0.11	0.14	0.24	0.53	0.86	0.97	.995
LW6	$P(1|\theta)$	0.21	0.22	0.29	0.58	0.895	0.98	1.00
	$P(1|(\theta + 2\sigma_\theta))$	0.31	0.31	0.35	0.63	0.925	0.99	1.00

were not found to be monotonically related to SOS values. The orderings of items given by SOS and P(SOS) are different, and this seems entirely proper. If an item is very poorly estimated it will probably have a large SOS value but a small P(SOS). For example, Linn et al.(1980, 1981) found a very large SOS value for their statistic for item 25. In this replication with their statistic and the present sample, that finding is confirmed by obtaining a very large SOS value. However, P(SOS) was very small. Examination of plots of confidence intervals showed the poor estimation.

For item 25, much of the discrepancy between the curves was for extreme values of θ where few subjects were available. Integration on the interval [-2.5, 1] instead of [-3, 3] resulted in a decreased SOS value. The change to the shorter ability interval gives higher weight to those segments of the curves having the most examinees for estimation. Linn has proposed defining an SOS statistic with $w(\theta)$ proportional to the number of examinees available for estimation at level θ. This statistic is currently under evaluation.

Limitations

The weakest link in our analysis is the estimation of the covariance matrices. In this author's opinion, LOGIST is clearly the best parameter estimation program available for the three-parameter logistic model. However, it has many ad hoc features (such as the assignment of default values), and its estimates only approximate maximum likelihood estimates. Moreover, its numerous options cause various deviations from maximum likelihood. Furthermore, the statistical theory for maximum likelihood estimation for item response models is incomplete. It has not been proven that maximum likelihood estimates are asymptotically normal with covariance matrices given by inverted information matrices. The assumption that covariance matrix entries are inversely proportional to sample size for fixed test length has not been proven. The method of estimating covariance matrices used in this paper ignores the error in estimating abilities. These problems are not insurmountable. An attempt to validate a method for computing covariance matrices for the joint estimation of abilities and item parameters is currently being undertaken. In addition there are promising developments in parameter estimation being completed in other laboratories. Such methods can be incorporated into the framework developed in this paper.

References

Durost, W.N., Bixler, H.H., Wrightstone, J.W., Prescott, G.A., and Balow, I.H. Metropolitan achievement tests, Form F. New York: Harcourt, Brace, and Jovanovich, 1970.

Kotz, S., Johnson, N. L., & Boyd, D. W. Series representations of distributions of quadratic forms in normal variables. I. Central case. Annals of Mathematical Statistics, 1967, 38, 823-837.

Levine, M.V. and Williams, B. A technique for calculating the distribution function for sums of squared normal random variables. ONR technical report in preparation (1982).

Linn, R.L., Levine, M.V., Hastings, C.N., and Wardrop, J. An investigation of item bias in a test of reading comprehension. Technical Report No. 163. Center for the Study of Reading, University of Illinois, Urbana. March 1980.

Linn, R.L., Levine, M.V., Hastings, C.N., and Wardrop, J. An investigation of item bias in a test of reading comprehension. Applied Psychological Measurement, 1981, 4, 159-173.

Lord, F. Applications of item response theory to practical testing problems. Hillsdale, New Jersey: Lawrence Earlbaum Associates. 1980.

Wood, R.L., Wingersky, M.S., and Lord, F.M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. (ETS RM 76-6). Princeton, New Jersey: Educational Testing Service, 1976.